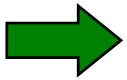
The JHU WS2006 IWSLT System

Experiments with Confusion Net Decoding

Wade Shen, Richard Zens, Nicola Bertoldi and Marcello Federico

Outline



- **Spoken Language Translation**
  - Motivations
  - ASR and MT
  - Statistical Approaches
- **Confusion Network Decoding**
  - Confusion Networks
  - Decoding of Confusion Network Input
  - Other Applications of Confusion Networks
- **Factored Models for TrueCasing**
- **Evaluation Experiments**

Motivations

Spoken Language Translation

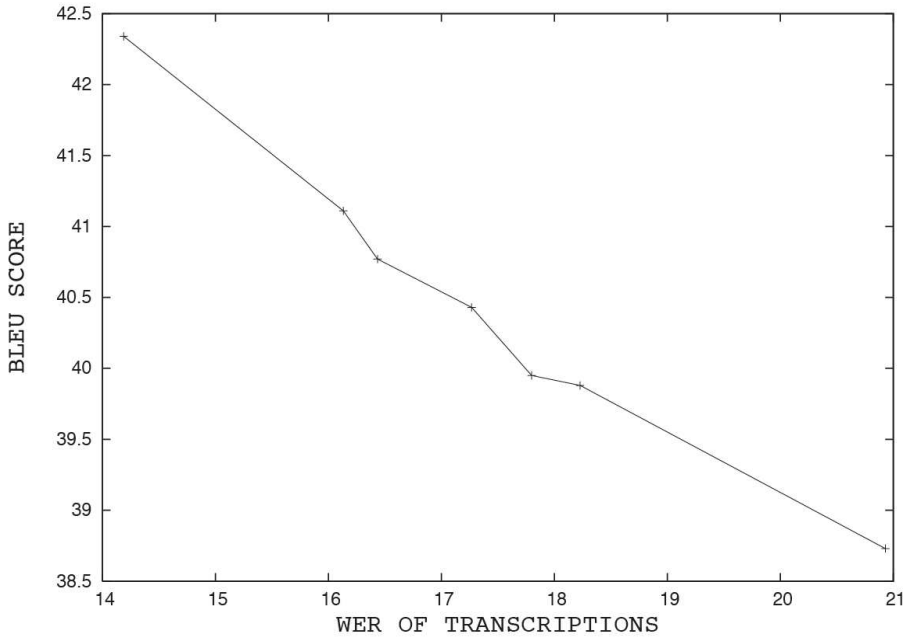
- **Translation from speech input is likely more difficult than translation from text**

- **Input**
  - **many styles and genres**
    formal read speech, unplanned speeches, interviews, spontaneous conversations, ...
  - **less controlled language**
    relaxed syntax, spontaneous speech phenomena
  - **automatic speech recognition is prone to errors**
    possible corruption of syntax and meaning

- **Need better integration for ASR and MT to improve spoken language translation**

Combining ASR and MT

- **Correlation between transcription word-error-rate and translation quality:**



- **Better transcriptions could have existed during ASR decoding: may get pruned for 1-best hypothesis**
- **Potential for improving translation quality by exploiting more transcription hypotheses generated during ASR.**

Spoken Language Translation

Statistical Approach

- Let $\mathbf{o}$ be the foreign language speech input
- Let $\mathcal{F}(\mathbf{o})$ be a set of possible transcriptions of $\mathbf{o}$

Goal – Find the best translation e* given this approximation:

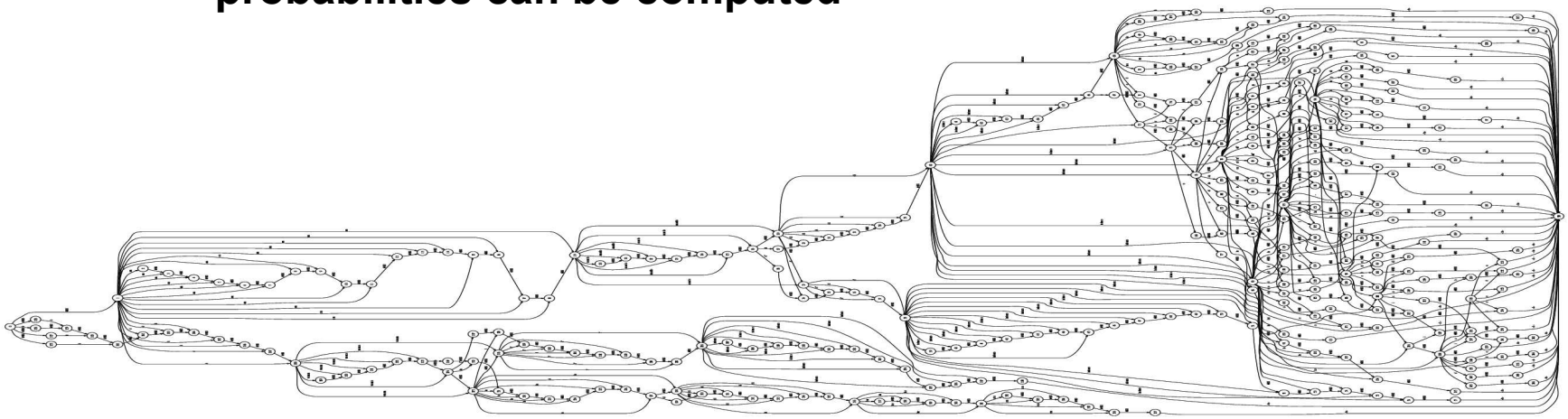
$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{o}) \approx \arg\max_{\mathbf{e}} \max_{f \in \mathcal{F}(\mathbf{o})} \Pr(\mathbf{e}, \mathbf{f}|\mathbf{o})$$

$\Pr(\mathbf{e}, \mathbf{f}|\mathbf{o})$ **is computed with a log-linear model with:**

- **Acoustics features:** *i.e. probs that some foreign words are in the input*
- **Linguistic features:** *i.e. probs of foreign and English sentences*
- **Translation features:** *i.e. probs of foreign phrases into English*
- **Alignment features:** *i.e. probs for word re-ordering*

ASR Word Graph

- **A very general set of transcriptions $\mathcal{F}(\mathrm{o})$ can be represented by a word-graph:**
  - **directly computed from the ASR word lattice (e.g. HTK format, `lattice-tool`)**
  - **provides a good representations of all hypotheses analyzed by the ASR system**
  - **arcs are labeled with words, acoustic and language model probabilities**
  - **paths correspond to transcription hypotheses for which probabilities can be computed**



Overview of SLT Approaches

- **1-best Translation:** *Translate most probable word-graph path*

Pros	**Most efficient**
Cons	no potential to recover from recognition errors

- **N-best Translation:** *Translate N most probable paths*

Pros	**Least efficient (linearly proportional to N)**
Cons	N must be large in order to include good transcriptions

- **Finite State Transducer:** *Compose WG with translation FSN*

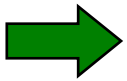
Pros	**Most straightforward, can examine full word graph**
Cons	Prohibitive with large vocabs and long range re-ordering

- **Confusion Network:** *translate linear approximation of WG*

Pros	**Can effectively explore graph w/o reordering problems**
Cons	Can overgenerate the input word graph

Outline

- Spoken Language Translation
  - Motivations
  - ASR and MT
  - Statistical Approaches



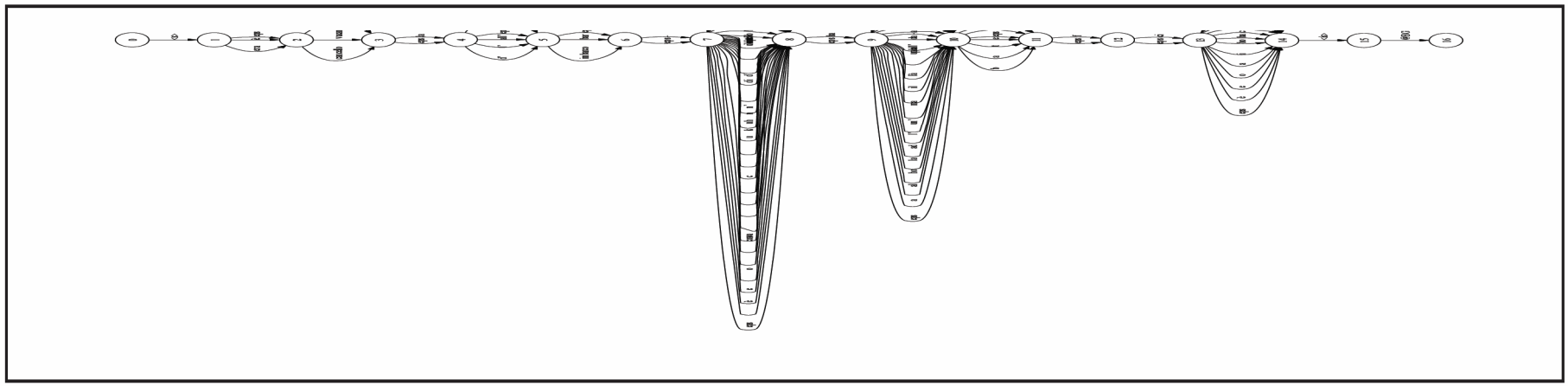
- Confusion Network Decoding
  - Confusion Networks
  - Decoding of Confusion Network Input
  - Other Applications of Confusion Networks

- Factored Models for TrueCasing

- Evaluation Experiments

Confusion Networks

- A **confusion network** approximates a word graph with a linear network, such that:
  - arcs are labeled with words or with the empty word (-word)
  - arcs are weighted with word posterior probabilities



- CNs can be conveniently represented as a sequence of *columns* of different depths

Confusion Network Decoding

Process

- **Extension of basic phrase-based decoding process:**
  - ***cover*** **some not yet covered consecutive columns (span)**
  - ***retrieve*** **phrase-translations for all paths inside the columns**
  - ***compute*** **translation, distortion and target language models**

- **Example: Coverage Vector = 01110…, path = cancello d'**

0	1	1	1	0	…
era $_{0.997}$ è $_{0.002}$ ϵ $_{0.001}$	cancello $_{0.995}$ vacanza $_{0.004}$ ϵ $_{0.002}$	ϵ $_{0.999}$ la $_{0.001}$	di $_{0.615}$ d' $_{0.376}$ all' $_{0.005}$ l' $_{0.002}$ ϵ $_{0.001}$	imbarco $_{0.999}$ bar $_{0.001}$	…

Confusion Net Decoding

Moses Implementation

- **Computational issues:**
  - **Number of paths grows exponentially with span length**
  - **Implies look-up of translations for a huge number of source phrases**
  - **Factored models require considering joint translation over all factors (tuples):**

    cartesian product of all translations of each single factor

- **Solutions implemented in Moses**
  - **Source entries of the phrase-table are stored with prefix-trees**
  - **Translations of all possible coverage sets are pre-fetched from disk**
  - **Efficiency achieved by incrementally pre-fetching over the span length**
  - **Phrase translations over all factors are extracted independently, then translation tuples are generated and pruned by adding a factor each time**

- **Once translation tuples are generated, usual decoding applies.**

Other Applications of Confusion Nets

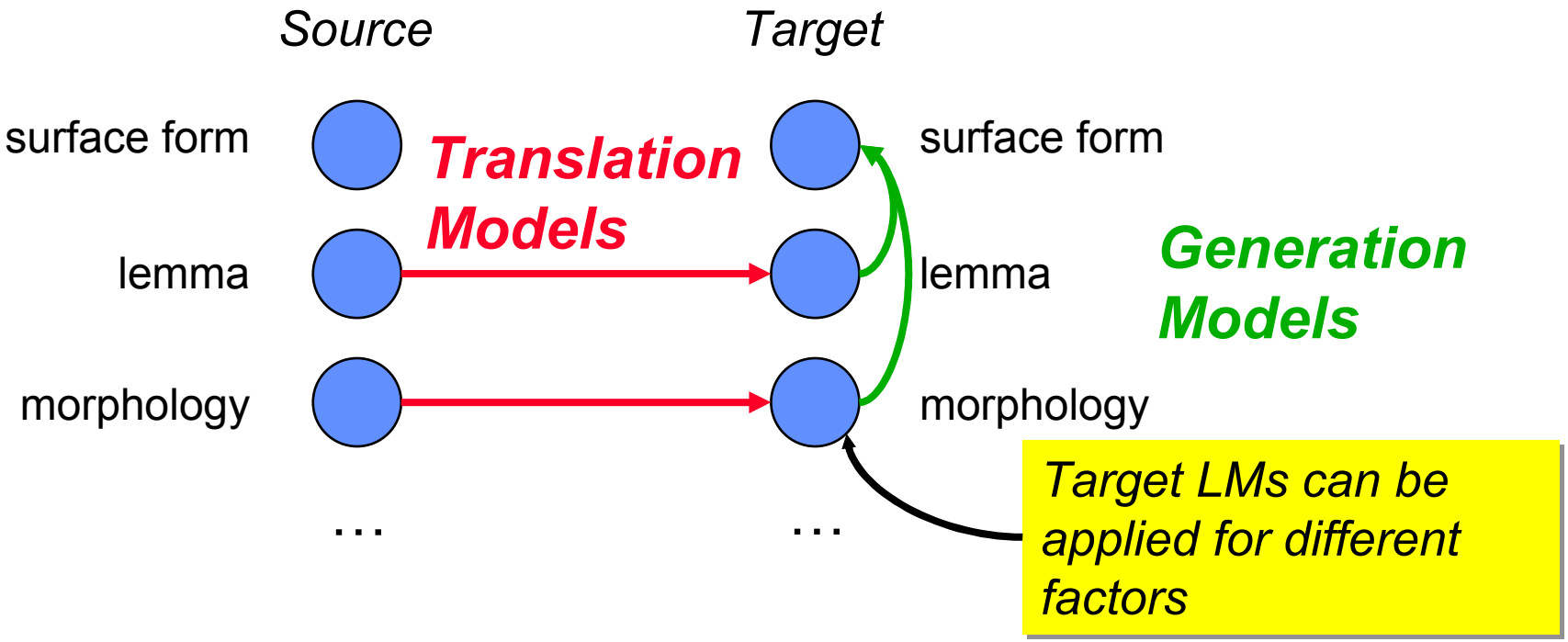
- Linguistic annotation for factored models
  - avoid hard decision by linguistic tools but rather provide alternative annotations with respective scores:
  - e.g. particularly ambiguous part of speech tags
- Translation of input similar to that produced by speech recognition
  - e.g. OCR output for optical text translation
- Insertion of punctuation marks missing in the input
  - model all possible insertions of punctuation marks in the input
- ...

Outline

- **Spoken Language Translation**
  - Motivations
  - ASR and MT
  - Statistical Approaches
- **Confusion Network Decoding**
  - Confusion Networks
  - Decoding of Confusion Network Input
  - Other Applications of Confusion Networks
- **Factored Models for TrueCasing**
- **Evaluation Experiments**

Factored Models

- **Factored representation**



- **Combine translation/generation/LMs in log-linear way**
- **Benefits**
  - **Generalization:** *Gather stats over generalized classes*
  - **Richer models:** *Can make use different linguistic representations*

Factored Models for TrueCasing

- Let $s_{1\ldots j}$ be the uncased word sequence
- Let $w_{1\ldots j}$ be the TrueCased word sequence

$$P(w_{1\ldots j}|s_{1\ldots j}) = \frac{P(s_{1\ldots j}|w_{1\ldots j}) * P(w_{1\ldots j})}{P(s_{1\ldots j})}$$

$$\arg\max_{w_{1\ldots j}} P(w_{1\ldots j}|s_{1\ldots j}) = \arg\max_{w_{1\ldots j}} P(s_{1\ldots j}|w_{1\ldots j}) * P(w_{1\ldots j})$$

$$\hat{P}(w_{1\ldots j}) \approx \prod_{k=1}^{j} P(w_k|w_{k-1}\ldots w_{k-n+1})$$ ***Mixed-case Language Model***

$$\hat{P}(s_{1\ldots j}|w_{1\ldots j}) \approx \prod_{k=1}^{j} P(s_k|w_k)$$ ***Generation Model***

- Translate lowercased, generate TrueCase, apply LM for both
  - Integrated into decoding
- Generation and language models jointly optimized with other translation models
  - Using Powell-like MER procedure

Outline

- **Spoken Language Translation**
  - Motivations
  - ASR and MT
  - Statistical Approaches
- **Confusion Network Decoding**
  - Confusion Networks
  - Decoding of Confusion Network Input
  - Other Applications of Confusion Networks
- **Factored Models for TrueCasing**
- **Evaluation Experiments**

Dev and Eval Corpus Statistics

- **Training Set Statistics (same models as MIT/LL)**

	Chinese	English
sentences	40 K	
running words	351 K	365 K
avg. sent. length	8.8	9.1
vocabulary entries	11 K	10 K

- **`Dev4` Confusion Network Statistics**

	speech type	
	read	spontaneous
avg. length	17.2	17.4
avg. / max. depth	2.2 / 92	2.9 / 82
avg. number of paths	10^{21}	10^{32}

- **`Dev4` and `test` Word Error Rates**

	speech type	
	read	spontaneous
`dev4`	12.8%	21.9%
`test`	15.2%	20.6%

Results

- **Overall Results**

test set	input	speech type: read BLEU [%]	speech type: spontaneous BLEU [%]
dev4	verbatim	21.4	
	1-best	19.0	17.2
	full CN	19.3	17.8
eval	verbatim	21.4	
	1-best	18.5	17.0
	full CN	18.6	18.1

- **Confusion Net Punctuation (dev4)**

punctuation input type	BLEU [%]
1-best	20.8
confusion network	21.0

- **Factored Truecasing (dev4)**

TrueCase Method	BLEU [%]
Standard Two-Pass: SMT + TrueCase	20.65
Integrated Factored Model (optimized)	21.08

Conclusions and Follow-on Work

- **Confusion net decoding shows significant gains**
  - Especially in spontaneous speech
  - Up to 6.4% relative improvement (higher WER?)
- **Confusion nets may be helpful for coupling MT with preprocessing steps**
  - Benefits with ASR
  - Modest benefits with repunctuation
- **Single pass TrueCasing may be helpful**
  - Joint decoding yields 2.0% relative increase
- **`moses` available (open source) for research**
  - http://www.statmt.org/moses/